HotRegion: a database of predicted hot spot clusters

Engin Cukuroglu, Attila Gursoy and Ozlem Keskin*

Center for Computational Biology and Bioinformatics and College of Engineering, Koc University, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey

ABSTRACT

Hot spots are energetically important residues at protein interfaces and they are not randomly distributed across the interface but rather clustered. These clustered hot spots form hot regions. Hot regions are important for the stability of protein complexes, as well as providing specificity to binding sites. We propose a database called HotRegion, which provides the hot region information of the interfaces by using predicted hot spot residues, and structural properties of these interface residues such as pair potentials of interface residues, accessible surface area (ASA) and relative ASA values of interface residues of both monomer and complex forms of proteins. Also, the 3D visualization of the interface and interactions among hot spot residues are provided. HotRegion is accessible at http://prism.ccbb.ku.edu.tr/ hotregion.

INTRODUCTION

Proteins interact with other proteins through their interfaces in order to fulfill their functions. Interfaces are formed by residues whose properties determine binding specificity and affinity. Correct orientations of the residues are critical for complex formation. Interactions between the residues in the binding sites are higher than the protein surface which shows that protein–protein interactions are highly depending on the cooperativity of the residues (1).

Some proteins interact with one or two proteins. Some other proteins, called hub proteins, may interact with many proteins as many as tens of other proteins. It is physically impossible for these hub proteins to interact with all its partners at the same time, since the surface area of the hub protein is fixed. This suggests that there are binding sites that should be used repeatedly to bind different proteins (2–4), probably each with different affinity and specificity. The distribution of the residues across the interface and the residue–residue interactions may answer the question 'How can the interfaces recognize their partners?'. The residues tend to behave cooperatively during the interactions and they form modules in the interface (5). Proteins may utilize these modules in order to have specificity and affinity during interactions (6–9) and also the combinations of these modules yield a powerful mechanism for binding multiple partners via unique interfaces (10,11). Also, Chakrabarti and Janin (12) stated that small binding sites have single continuous patch; however, larger interfaces may have several patches. Previously, modules in interfaces are defined with various methods such as (i) the edge betweenness criteria in the residue–residue interaction network across the interface (7,13), (ii) difference of energy profiles of residues in interfaces (6,14,15), and (iii) clustering of structurally conserved residues in interfaces (9,11,16). In the edge betweenness approach, the authors used the topology of the network without considering residue energy profiles. The other two approaches used hot spot residues which are driven by energy profiles or structural conservation of residues. The residues that contribute more to the binding free energy are called 'hot spots' (17–19). Hot spots are tightly packed and structurally conserved residues (9,11,16). Also Keskin *et al.* (9) showed that these hot spot residues are not randomly distributed along the protein–protein interfaces; rather clustered. Besides, there is a correlation between energy change and decrease in the accessible surface area of these hot spots (20). Also, the cooperativity of the hot spot residues enlightens the complex binding organizations of the protein–protein interfaces (21,22,23). Computational methods (24–30) are widely used to extract hot spot information from interface, because experimental studies are available for a very limited number of complexes.

In this work, we combine the residue network topology with the residue energy profile based clustering approaches. The residue clusters in interfaces are called 'hot regions' (9,22). As we showed in our previous study

*To whom correspondence should be addressed. Tel: +90 212 338 1538; Fax: +90 212 338 1548; Email: okeskin@ku.edu.tr
Correspondence may also be addressed to Attila Gursoy. Tel: +90 212 338 1720; Fax: +90 212 338 1548; Email: agursoy@ku.edu.tr

(22), hot regions are useful to interpret the protein interface properties. Here, we present our database 'HotRegion' in order to illustrate hot spot cooperativity information at protein–protein interfaces. HotRegion stores all available protein–protein interfaces which are extracted from Protein Data Bank (PDB) (31) entries using a dynamic update system which is based on the user's search queries. If a user searches for hot regions via a PDB ID which is not in the HotRegion database, the database can rapidly update itself and show the results. We hope the database will help in detecting cooperativity of functionally important residues, mutagenesis targets and understand the stability and specificity of protein–protein interfaces.

HOTREGION METHOD

An interface is the contact region between two interacting proteins. Two residues are defined to be contacting if the distance between any two atoms of the two residues from different chains is less than the sum of their corresponding van der Waals radii plus 0.5 Å (32,33). Hotspot residues in interfaces are predicted with HotPoint (28) using accessible surface area (ASA) and knowledge based pair energies of each residue (34). In order to define hot regions, a network of hotspots is constructed. In the network, the nodes are the hotspot residues and the edges are linked between nodes if the two hotspot residues are in contact. Two hotspot residues are defined as contacting if the distance between their $C\alpha$ atoms is smaller than 6.5 Å (9). Afterwards, connected components of the network are found and if the nodes in a connected component are equal or greater than three, the connected component is labeled as a hot region and the hotspot residues in this connected component labeled as the members of this hot region.

DATABASE PROPERTIES

The HotRegion database is available at http://prism.ccbb .ku.edu.tr/hotregion. HotRegion embraces three major components: a relational database management system for data storage and management, a web application to interface the database and a dynamically database update system. Data are stored in a relational MySQL database. The web application runs on an Apache web server hosted on a linux-based system. PHP and JavaScript are used to implement the web application. The database can be updated dynamically.

DATABASE CONTENT

Currently, HotRegion contains all the PDB entries as of January 2011 (70 695 PDB entries, 147 892 protein–protein interfaces). If a user searches hot region information of a protein complex (via PDB ID) which is not in the HotRegion, the database can rapidly update itself and show the results. HotRegion has only protein–protein interface information. HotRegion database offers the researchers to find the hot regions of the protein complexes and provides structural properties of these complexes such as pair potentials of interface residues, ASA and relative ASA values of interface residues of both monomer and complex forms of proteins. Also, the visualization of the interface by using Jmol (35) and residue networks of interactions of hot spot residues are presented in the results. An advanced search option is also available. Users can manipulate the HotRegion parameters by changing default values in advanced search section. Advanced searches are deposited in the database and users can retrieve their jobs by using an email and job id from the 'Retrieve Job' section.

HotRegion needs atomic coordinates of the protein complexes in the standard PDB format. If atoms are present in alternative locations, only the first location is considered. For NMR structures, the first model is used. HotRegion is specific to protein–protein interfaces; chains corresponding to DNA and RNA structures return no interface solutions.

If users do not supply enough information, the database asks for the missing information. The HotRegion database is free, open to all users and there are no login requirements.

TUTORIAL

Simple search

Users retrieve the data of protein interfaces just by entering a PDB ID and two chain identifiers. Between the given monomers there must be an interface in order to get the hot region information. Also users have a control over the presentation of the results. Three properties of the interface (residue number, residue type, chain id) are always displayed in the result table and the output file, and the rest are displayed based on the preferences (Figure 1).

Advanced search

Users can retrieve the data based on their interface and hot region finding criteria. Users must enter email information in order to retrieve their jobs afterwards. They can supply a PDB file or enter a PDB code. After entering the chain information of the monomers that have an interface between them, users can decide a valid interface extraction threshold which is summed with van der Waals radii of atoms. When the van der Waals threshold gets bigger, the number of interface residues will increase. Also users can change the hot spot neighbor criterion which is the $C\alpha$ distance between the hot spots. When the hot region criterion gets bigger, the number of hot regions will decrease and hot regions start to merge in order to build larger hot regions.

Retrieve Job

The returning users can retrieve results of previous jobs by using the job ids and their email addresses.
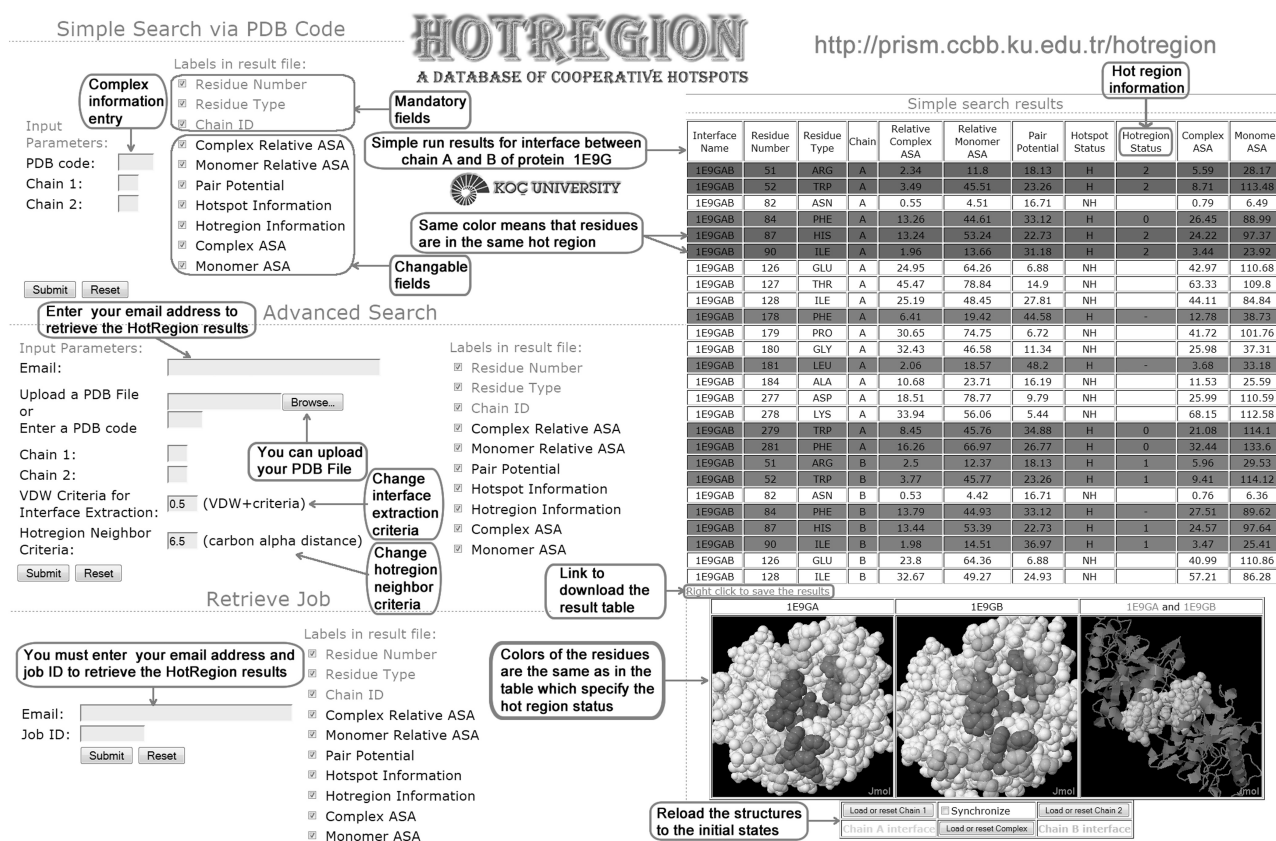
http://prism.ccbb.ku.edu.tr/hotregion

Interface Name	Residue Number	Residue Type	Chain	Relative Complex ASA	Relative Monomer ASA	Pair Potential	Hotspot Status	Hotregion Status	Complex ASA	Monomer ASA
1E9GAB	51	ARG	A	2.34	11.8	18.13	H	2	5.59	28.17
1E9GAB	52	TRP	A	3.49	45.51	23.26	H	2	8.71	113.48
1E9GAB	82	ASN	A	0.55	4.51	16.71	NH		0.79	6.49
1E9GAB	84	PHE	A	13.26	44.61	33.12	H	0	26.45	88.99
1E9GAB	87	HIS	A	13.24	53.24	22.73	H	2	24.22	97.37
1E9GAB	90	ILE	A	1.96	13.66	31.18	H	2	3.44	23.92
1E9GAB	126	GLU	A	24.95	64.26	6.88	NH		42.97	110.68
1E9GAB	127	THR	A	45.47	78.84	14.9	NH		63.33	109.8
1E9GAB	128	ILE	A	25.19	48.45	27.81	NH		44.11	84.84
1E9GAB	178	PHE	A	6.41	19.42	44.58	H	-	12.78	38.73
1E9GAB	179	PRO	A	30.65	74.75	6.72	NH		41.72	101.76
1E9GAB	180	GLY	A	32.43	46.58	11.34	NH		25.98	37.31
1E9GAB	181	LEU	A	2.06	18.57	48.2	H	-	3.68	33.18
1E9GAB	184	ALA	A	10.68	23.71	16.19	NH		11.53	25.59
1E9GAB	277	ASP	A	18.51	78.77	9.79	NH		25.99	110.59
1E9GAB	278	LYS	A	33.94	56.06	5.44	NH		68.15	112.58
1E9GAB	279	TRP	A	8.45	45.76	34.88	H	0	21.08	114.1
1E9GAB	281	PHE	A	16.26	66.97	26.77	H	0	32.44	133.6
1E9GAB	51	ARG	B	2.5	12.37	18.13	H	1	5.96	29.53
1E9GAB	52	TRP	B	3.77	45.77	23.26	H	1	9.41	114.12
1E9GAB	82	ASN	B	0.53	4.42	16.71	NH		0.76	6.36
1E9GAB	84	PHE	B	13.79	44.93	33.12	H	-	27.51	89.62
1E9GAB	87	HIS	B	13.44	53.39	22.73	H	1	24.57	97.64
1E9GAB	90	ILE	B	1.98	14.51	36.97	H	1	3.47	25.41
1E9GAB	126	GLU	B	23.8	64.36	6.88	NH		40.99	110.86
1E9GAB	128	ILE	B	32.67	49.27	24.93	NH		57.21	86.28

Figure 1. Properties of HotRegion Database in a quick view.

CASE STUDY

Contribution to binding affinity of the proteins

Colicins are plasmid-encoded, stress-induced protein antibiotics that specifically target *Escherichia coli* cells. When it binds to a specific (cognate) partner, the nuclease can protect the organism from endogenous and incoming colicin (36). Kleanthous and coworkers (37) showed that a limited number of mutations at the interface provide high-affinity binding to a non-cognate partner. According to this work, a non-cognate complex between the colicin E9 endonuclease (E9 DNase) and immunity protein 2 (Im2) (PDB Id: 2WPT) has a weaker binding affinity than the cognate femtomolar E9 DNase—Im9 (PDB Id: 1EMV) interaction. When they substitute three Im2 residues with their Im9 counterparts (Im2 D33L/N34V/R38T) the binding energy is almost similar to the binding energy of cognate complex energy. HotRegion results for these complexes show that the predicted hot spots overlap with the experimental findings. The cognate complex has two hot regions but the non-cognate complex has one hot region (Figure 2) (Table 1). The structural differences at the interface are based on the different side chain orientations. Possibly, cognate complex utilizes the two hot regions at the interface in order to increase the binding affinity of interaction. When we compare the hot regions of both complexes, we observed that the only difference between the hot region residues at
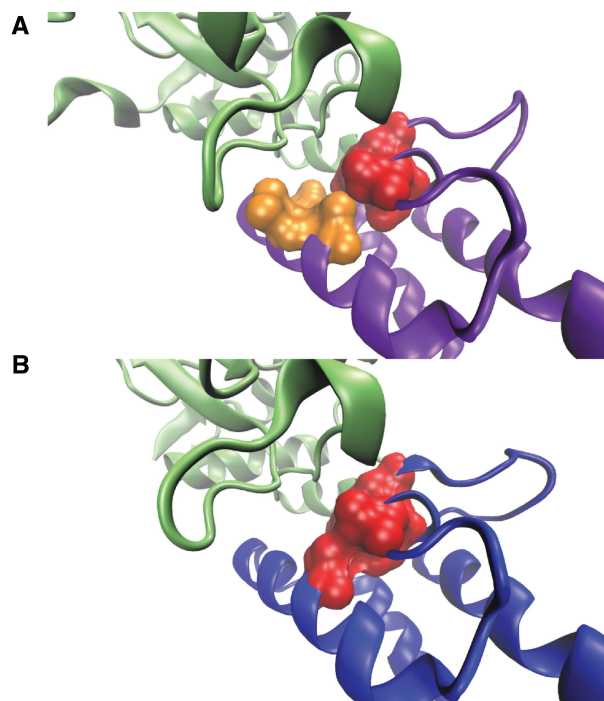
Figure 2. (A) Colicin E9 endonuclease (green) interacts with Im9 (purple) and the complex has two hot regions (red and orange). (B) Colicin E9 endonuclease (green) interacts with Im2 (blue) and the complex has one hot region (red).

Table 1. Hot region results from HotRegion Database for interfaces 1EMVAB and 2WPTAB

Interface	Residue number	Residue type	Chain	Hot region identifier
1EMVAB	33	LEU	A	1
1EMVAB	34	VAL	A	1
1EMVAB	37	VAL	A	1
1EMVAB	50	SER	A	0
1EMVAB	53	ILE	A	0
1EMVAB	54	TYR	A	0
2WPTAB	37	VAL	A	0
2WPTAB	50	SER	A	0
2WPTAB	53	ILE	A	0
2WPTAB	54	TYR	A	0

the cognate complex are L33 and V34 (they formed the extra hot region with T37 in cognate complex). When these residues are mutated in the non-cognate complex to L and V, these residues may probably form the extra hot region with T37 at non-cognate complex in order to increase the binding affinity of the non-cognate complex as much as the one of the cognate complex.

CONCLUSION

A protein–protein interface consists of two binding sites of two proteins interacting with each other. For all different protein interactions, the binding energies of each complex are miscellaneous and the hot spot residues are distributed in a distinctive pattern. Extracting hot region information from not uniformly distributed binding energy of interfaces is important for analyzing the binding sites of the proteins. Some complexes are built upon more than one hot region, and size of the hot region is changing according to the binding site properties.

We have earlier shown that such hot regions (hotspot clusters) are a signature for the protein–protein interfaces especially for hub proteins (22). A hub protein binds different partner proteins by using different hot regions. These networked hotspot organization may imply that the contribution of the hotspots to the stability of the protein–protein complex within a hot region is cooperative. We hope the database will help in detecting cooperativity of functionally important residues, mutagenesis targets and understand the stability and specificity of protein–protein interfaces.

Conflict of interest statement. None declared.

REFERENCES

1. Illingworth,C.J., Scott,P.D., Parkes,K.E., Snell,C.R., Campbell,M.P. and Reynolds,C.A. (2010) Connectivity and binding-site recognition: applications relevant to drug design. *J. Comput. Chem.*, **31**, 2677–2688.
2. Kim,P.M., Lu,L.J., Xia,Y. and Gerstein,M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.
3. Keskin,O. and Nussinov,R. (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**, 341–354.
4. Tuncbag,N., Gursoy,A., Guney,E., Nussinov,R. and Keskin,O. (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.
5. Liu,T., Whitten,S.T. and Hilser,V.J. (2007) Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proc. Natl Acad. Sci. USA*, **104**, 4347–4352.
6. Reichmann,D., Rahat,O., Albeck,S., Meged,R., Dym,O. and Schreiber,G. (2005) The modular architecture of protein-protein binding interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 57–62.
7. Del Sol,A., Arauzo-Bravo,M.J., Amoros,D. and Nussinov,R. (2007) Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biol.*, **8**, R92.
8. Tyagi,M., Shoemaker,B.A., Bryant,S.H. and Panchenko,A.R. (2009) Exploring functional roles of multibinding protein interfaces. *Protein Sci.*, **18**, 1674–1683.
9. Keskin,O., Ma,B. and Nussinov,R. (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.
10. Martin,J. (2010) Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. *PLoS Comput. Biol.*, **6**, e1000821.
11. Keskin,O., Ma,B., Rogale,K., Gunasekaran,K. and Nussinov,R. (2005) Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.*, **2**, S24–S35.
12. Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.
13. Carbonell,P., Nussinov,R. and del Sol,A. (2009) Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics*, **9**, 1744–1753.
14. Humphris,E.L. and Kortemme,T. (2007) Design of multi-specificity in protein interfaces. *PLoS Comput. Biol.*, **3**, e164.
15. Moza,B., Buonpane,R.A., Zhu,P., Herfst,C.A., Rahman,A.K., McCormick,J.K., Kranz,D.M. and Sundberg,E.J. (2006) Long-range cooperative binding effects in a T cell receptor variable domain. *Proc. Natl Acad. Sci. USA*, **103**, 9867–9872.
16. Halperin,I., Wolfson,H. and Nussinov,R. (2004) Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure*, **12**, 1027–1038.
17. Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.
18. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
19. Wells,J.A. (1991) Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol.*, **202**, 390–411.

20. Guharoy,M. and Chakrabarti,P. (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.

21. Shulman-Peleg,A., Shatsky,M., Nussinov,R. and Wolfson,H.J. (2007) Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.*, **5**, 43.

22. Cukuroglu,E., Gursoy,A. and Keskin,O. (2010) Analysis of hot region organization in hub proteins. *Ann. Biomed. Eng.*, **38**, 2068–2078.

23. Ahmad,S., Keskin,O., Mizuguchi,K., Sarai,A. and Nussinov,R. (2010) CCRXP: exploring clusters of conserved residues in protein structures. *Nucleic Acids Res.*, **38**, W398–W401.

24. Armon,A., Graur,D. and Ben-Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.

25. Guharoy,M. and Chakrabarti,P. (2009) Empirical estimation of the energetic contribution of individual interface residues in structures of protein-protein complexes. *J. Comput. Aided Mol. Des.*, **23**, 645–654.

26. Guney,E., Tuncbag,N., Keskin,O. and Gursoy,A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.

27. Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.

28. Tuncbag,N., Keskin,O. and Gursoy,A. (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–W406.

29. Zhu,X. and Mitchell,J.C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.

30. Shulman-Peleg,A., Shatsky,M., Nussinov,R. and Wolfson,H.J. (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.*, **36**, W260–W264.

31. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

32. Keskin,O., Tsai,C.J., Wolfson,H. and Nussinov,R. (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

33. Tsai,C.J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.

34. Tuncbag,N., Gursoy,A. and Keskin,O. (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

35. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.

36. Kleanthous,C. and Walker,D. (2001) Immunity proteins: enzyme inhibitors that avoid the active site. *Trends Biochem. Sci.*, **26**, 624–631.

37. Meenan,N.A., Sharma,A., Fleishman,S.J., Macdonald,C.J., Morel,B., Boetzel,R., Moore,G.R., Baker,D. and Kleanthous,C. (2010) The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc. Natl Acad. Sci. USA*, **107**, 10080–10085.